Abstract Title Page
Not included in page count.

Title: A Powerful, Potential Outcomes Method for Estimating Any Estimand across Multiple Groups

Authors and Affiliations:

Cassandra W. Pattanayak, Harvard University Statistics Department, pattanayak@stat.harvard.edu

Donald B. Rubin, Harvard University Statistics Department, rubin@stat.harvard.edu

Elizabeth R. Zell, Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, ezr1@cdc.gov

Abstract Body.

Context:

In educational research, outcome measures are often estimated across separate studies or across schools, districts, or other subgroups to assess the overall causal effect of an active treatment versus a control treatment. Students may be partitioned into such strata or blocks by experimental design, or separated into studies within a meta-analysis. In non-randomized studies, students may be partitioned into subclasses based on key covariates or estimated propensity scores to improve observed covariate balance across treatment groups (e.g., Rosenbaum & Rubin, 1983).

Procedures designed to estimate any estimand in the presence of strata, including a simple t-test for the difference in mean outcomes (Neyman, Iwaszkiewicz, & Kolodziejczyk, 1935), rely on implicit assumptions about the unknowable correlation between potential outcomes under active treatment and control treatment. For binary outcomes, the standard procedures used to estimate overall odds ratios in the presence of strata were introduced by Cochran (1954), who first proposed a hypothesis test for the difference in proportions across strata. Mantel and Haenszel (1959) proposed a very similar test and introduced an estimator for a common odds ratio.

Objective:

Consider the following hypothetical studies designed to estimate the causal effect of an existing program on high school graduation:

1. Within each of several school districts, half of the schools are randomized to participate in the program, and half are randomized not to participate in the program.
2. Within each of several cities, schools participating in the program are compared to schools not participating in the program, though participation was not randomized.
3. Several separate evaluations of the program are collected, to be combined in a meta-analysis.

In each hypothetical study, a binary outcome (graduation) must be measured over strata (1. school districts, 2. cities, 3. evaluations). The effect of the program on graduation rates may be measured by a difference in proportions, odds ratio, or some other quantity. A hypothesis test will be conducted and a confidence interval constructed for the chosen estimand.

We propose tests and intervals that can be more powerful for any finite population estimand than traditional tests and intervals, including t-tests for the difference in means or Cochran-Mantel-Haenszel procedures for the odds ratio.

Significance:

We show that the asymptotic sampling variance estimators typically used for the point estimates in the presence of strata (Neyman, Iwaszkiewicz, & Kolodziejczyk, 1935; Robins, Breslow, & Greenland, 1986; Robins, Greenland, & Breslow, 1986) can lead to unnecessarily wide confidence intervals and weak hypothesis tests. We propose a Bayesian approach that explicitly

imputes missing potential outcomes under the Rubin Causal Model (Holland, 1986; Rubin, 1974; Rubin, 1978). We demonstrate, by a simulation study, that in many circumstances the proposed estimator has greater precision and leads to more accurate interval coverage rates than traditional procedures. The proposed method also avoids homogeneity assumptions; stratified study designs are motivated by an assumption that the treatment effects vary across strata, yet standard procedures typically assume homogeneous treatment effects.

Unlike traditional methods, our procedure does not rely on assumptions about asymptotic normality that may not be met in small samples, and our procedure leads to confidence intervals that have Bayesian interpretation in addition to Frequentist coverage.

We also clarify that too-wide intervals that err on the side of failing to reject the null hypothesis are often not "conservative": a cautious researcher considering whether to cancel an academic program would err on the side of stating that the program does have a positive effect.

Statistical Model:

Consider a finite population of N students, with N_b students in each stratum b $(b = 1,\ldots,B)$, where $N_1+\ldots+N_B=N$. Each student i $(i = 1,\ldots,N)$ has a fixed potential outcome under treatment, $Y_i(1)$, and a fixed potential outcome under control, $Y_i(0)$, only one of which can be realized and observed. We assume (1) that the outcome Y is binary, such as a graduation indicator, and (2) that the stable unit treatment value assumption (SUTVA) holds (Rubin, 1980).

If $Y_i(1)$ and $Y_i(0)$ were both known for each student, then the finite population difference in means, odds ratio, or any other estimand could be *calculated* rather than *estimated*. The fundamental challenge of causal inference is that at most one potential outcome can be observed for each student. Following Rubin (1978), we propose a Bayesian model for explicitly imputing the missing potential outcome for each student. We impute the missing potential outcomes using binomial models within each stratum, independently imputing the active treatment and control potential outcomes. By imputing all of the missing potential outcomes multiple times and calculating the estimand for each imputation, we can generate a point estimate and posterior interval for any estimand. This model-based imputation approach differs from other model-based methods in that the estimand is not forced to be a parameter in some super-population model.

Research Design:

We conducted a simulation study to compare the proposed methods with traditional procedures. The simulation parameters (sample size, number of subclasses, outcome rates, and heterogeneity of rates across strata) were set to reflect conditions that commonly arise when propensity score subclasses are created to address covariate imbalances in randomized or non-randomized studies. However, the conditions examined in this simulation also arise in randomized block experiments, meta-analyses, and other situations.

Results:

The substantive results of the simulation study did not depend on the simulation parameters; therefore, tables of results would be repetitive, and our conclusions can be briefly summarized.

Our method and the traditional estimators generate approximately equal mean absolute percent biases, though Cochran-Mantel-Haenszel estimates can be more biased than our odds ratio estimates in the presence of heterogeneous strata. However, traditional t-tests and the Cochran-Mantel Haenszel procedure generate intervals approximately 1.4 times wider than the intervals produced by our method.

Usefulness / Applicability of Method:

Suppose that in one of the hypothetical studies described above, the graduation rate was 0.6 among students enrolled in the program of interest and 0.5 among students not in the program. An estimate for the average causal effect of the program on graduation rates is 0.1. If the 95% confidence interval generated by a t-test was (-.02, .22), the corresponding interval from our procedure might be (0.02, 0.19). Our method errs on the side of detecting an effect, while traditional methods err on the side of failing to detect an effect. A preference for either method depends on the situation: if we are considering eliminating a program that took a long time to develop, for example, it would be more cautious to over-estimate rather than under-estimate the program's effect, minimizing the chances that an effective program would be eliminated.

Conclusions:

We propose a Bayesian, multiple imputation procedure for estimating any finite population estimand in the presence of subclasses. Intervals from standard t-tests or Cochran-Mantel-Haenszel procedures are approximately 1.4 times as wide as the corresponding intervals generated from our method. Therefore, our method leads to hypothesis tests that are more powerful than standard procedures.

The difference in efficiency between traditional procedures and the method we propose is due to differing assumptions about the unobservable correlation, conditional on strata, between (a) potential outcomes under active treatment and (b) potential outcomes under control treatment. The assumption of perfect correlation between active treatment and control potential outcomes underlies the Cochran-Mantel-Haenszel procedures and standard Neyman confidence intervals (Neyman, Iwaszkiewicz, & Kolodziejczyk, 1935) which tend to cover the true causal effect in finite samples at least as often as the nominal coverage. Hypothesis tests based on standard Neyman asymptotic confidence intervals are defined to be statistically "conservative" because the Type I error tends to be less than or equal to the nominal level in a finite sample. But when the goal is to identify small differences between two treatments, a cautious hypothesis test would err on the side of rejecting the null hypothesis too frequently, indicating a possible difference, rather than rejecting too infrequently. The intervals and corresponding hypothesis tests that we propose minimize the Type II error rate, increasing the power of the tests to detect non-null values of the estimand. These considerations, and generalizations of the proposed method, apply to any statistic or study design.

In many real data sets with strata based on relevant covariates, we believe that the conditional correlation between potential outcomes may be closer to zero than to one, suggesting that our narrower intervals are more appropriate than traditional ones. This is especially true when the strata are based on covariates highly predictive of treatment decisions and outcomes.

When the estimand is defined in terms of a broader population from which the observed data is thought to be a random sample, the corresponding proposed procedures impute all missing potential outcomes: one unobserved potential outcome for each unit included in the sample, and two unobserved potential outcomes for all units not included in the sample. These super-population methods lead to intervals with bias, width, and coverage similar to traditional methods, at least in modest to large samples, which is expected because traditional intervals are designed to over-cover for the finite population estimand and cover at the nominal rate for the super-population estimand, at least asymptotically. However, our methods do not depend on asymptotic normality.

Appendices

Appendix A. References

Cochran, W. (1954). Some methods for strengthening the common chi-square tests. *Biometrics, 10*(4), 417-451.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Neyman, J., Iwaszkiewicz, K., & Kolodziejczyk, S.1935. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society, 2*, 107-154.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42*(2), 311-323.

Robins, J., Greenland, S., & Breslow, N. (1986). A general estimator for the variance of the Mantel-Haenszel odds ratio. *American Journal of Epidemiology, 124*(5), 719-723.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6*(1), 34-58.

Rubin, D. B. (1980). Discussion of 'Randomization analysis of experimental data in the Fisher randomization test' by Basu.' *Journal of the American Statistical Association, 75*(371), 591-593.